多階層アラインメント視覚表現に基づくリアルタイム物体操作タスク成功判定

Real-Time Task Success Prediction for Open-Vocabulary Manipulation Based on Muli-Level Aligned Representation

後神 美結 Miyu Goko　神原 元就 Motonari Kambara　齋藤 大地 Daichi Saito　小槻 誠太郎 Seitaro Otsuki　杉浦 孔明 Komei Sugiura

慶應義塾大学
Keio University

Task success prediction is important for open-vocabulary object manipulation tasks by manipulators, because it can enhance the reliability and efficiency of manipulations. In particular, it is highly convenient for efficient task execution if success prediction can be performed on-the-fly during the object manipulation. We propose a framework that extends Contrastive λ-Repformer, which determines task success based on images before and after open-vocabulary object manipulation and a instruction, to real-time task success prediction for open-vocabulary manipulation. Our framework detects the timing of successful object manipulation by focusing on minute changes between images, which is conducted by contrasting multi-level aligned representation for images in the initial state and images at any given time. Experimental results confirm that real-time success prediction could be performed using this framework.

1. はじめに

物体操作におけるタスク成功判定は，操作の正確性と効率性を高めることができ，医療，産業，農業，物流を含む業界でのロボット応用において，信頼性と一貫性の向上に寄与する．例えば，産業における部品の組み立てや，農業における作物の収穫などの物体操作タスクにおいて，マニピュレータがタスクの成功を判定できれば，品質や生産性の向上が期待できる．また，サブタスクの失敗が後続のタスクに影響を及ぼす可能性があるため，サブタスクごとの成否を正確に判定することは，長期的なタスク遂行において特に重要である．

本タスクは，以下の 2 点により困難である．まず，物体操作前後に撮影された画像の変化，画像内の物体に関する情報，open-vocabulary 物体操作の指示文を十分に理解する必要がある点である．さらに，本タスクでは，上記の要素が互いに適切にアラインしているかを判断することが求められる．既存のマルチモーダル大規模言語モデル（MLLM [Achiam 23,GeminiTeam 23,Dai 23]）を本タスクに単純に適用した場合，その性能は限定的である．これは，MLLM が物体の詳細な特徴や位置の微細な変化を十分に捉えることが難しいためである．

本研究では，画像と指示文のアラインメントを行うことで，open-vocabulary 物体操作タスクの成否を予測する Contrastive λ-Repformer を提案する．本手法は，我々の先行研究である λ-Repformer [齋藤 24] における視覚表現 λ-Representation を拡張し，活用する．λ-Representation は，(i) 局所的な画像情報を保持する特徴，(ii) 自然言語とアラインされた特徴，(iii) 自然言語によって構造化された特徴の 3 種類の要素を統合した視覚表現である．従来手法では，単一の視覚表現抽出メカニズムに依存しているため，物体のテクスチャや形状といった詳細な視覚特徴と，物体間の空間的関係といったグローバルな構造表現の両方を十分に捉えることが難しいという課題があった．さらに，提案手法では，画像間の差異を表現する特徴量を用いるため，指示文とこの差異を効果的にアラインすることができる．これにより，本手法では物体の詳細な特徴や，それらの空間的関係を考慮した指示文の理解が可能となる．

本研究の新規性は以下の通りである．

- 画像に対して前述の 3 種類の視覚表現を抽出し，λ-Representation として統合する λ-Representation Encoder を導入する．
- 2 つの画像の λ-Representation の差異を抽出する Contrastive λ-Representation Decoder を提案する．このモジュールを用いることで，タスク成功予測の際に画像間の変化と指示文とのアラインメントを考慮することが可能となる．

連絡先: 後神美結，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，miyu.goko@keio.jp

2. 問題設定

本研究では，Success Prediction for Open-vocabulary Manipulation（SPOM）タスクを扱う．SPOM とは，物体操作の前後に撮影された一人称視点画像と指示文を入力とし，open-vocabulary 物体操作タスクの成否を予測するタスクである．本タスクでは，モデルが物体操作の成功または失敗を適切に判定することが求められる．

入力は，物体操作前後に撮影された一人称視点画像と指示文で構成される．出力は，指示文に基づく物体操作が適切に実行された確率の予測値を示す予測確率 $P(\hat{y}=1)$ である．ここで，$\hat{y}$ は物体操作の成否を示す予測ラベルであり，操作に成功していると判定されるとき '1' である．本研究では，一人称視点画像のみを入力画像として使用する．また，一部の画像では，マニピュレータによりシーンの一部が遮られている．このとき，タスク自体は実行可能であるが，対象物やエリアが部分的に隠れることで，遂行が困難となるサンプルも含まれる．

3. 提案手法

図 1 に提案手法である Contrastive λ-Repformer の構造を示す．入力は $\boldsymbol{x}=\{\boldsymbol{x}_{\text{inst}},\boldsymbol{x}_{\text{before}},\boldsymbol{x}_{\text{after}}\}$ であり，ここで $\boldsymbol{x}_{\text{inst}}$ はトークン化された指示文，$\boldsymbol{x}_{\text{before}}$ および $\boldsymbol{x}_{\text{after}}$ は，それぞれ物体操作前後に撮られた RGB 画像を表す．提案手法の主要なモジュールは λ-Representation Encoder と Contrastive λ-Representation Decoder である．本研究は Contrastive λ-Repformer をリアルタイム物体操作タスク成功判定へと応用

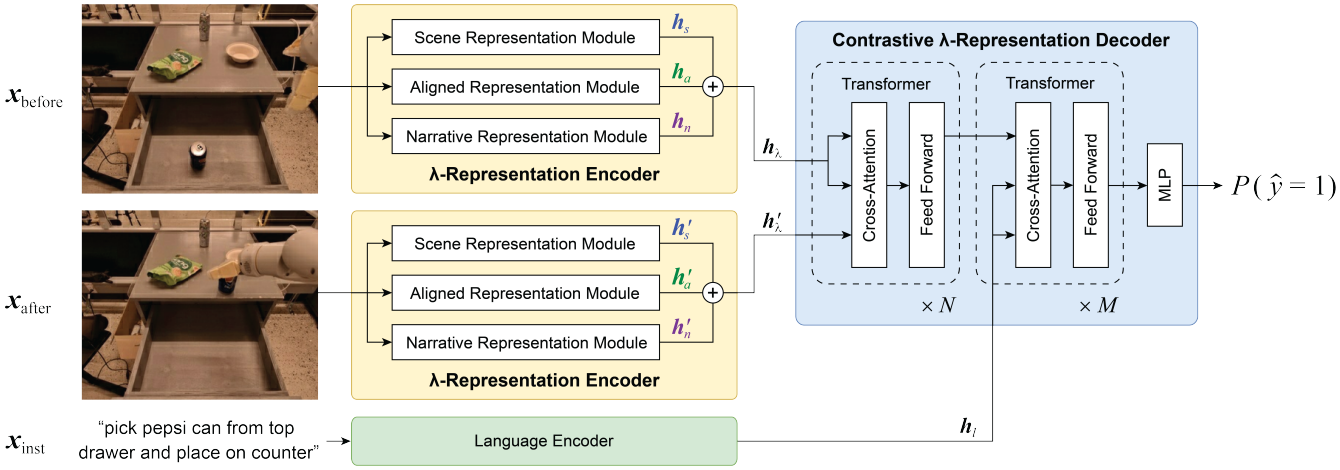

図 1: Contrastive λ-Repformer の概要．指示文と物体操作前後の画像を入力として，タスク成功確率の推定値を出力する．

しており，各モジュールの詳細については [Goko 24] に記載されている．

3.1 λ-Representation

既存の Vision-and-Language 研究において，視覚特徴の抽出には主に三つのアプローチが存在する．1 つのアプローチは，ユニモーダル画像エンコーダ [Dosovitskiy 20, Liu 21, Darcet 24] を用いるものであり，テクスチャやエッジといった視覚特徴を抽出する．本研究では，これらの特徴を Scene Representation と定義する．次のアプローチは，マルチモーダル画像エンコーダ [Radford 21, Chen 20, Li 22, Zhai 23] を活用し，自然言語とアラインされた視覚特徴を抽出するものである．これらの特徴を Aligned Representation と呼ぶ．最後のアプローチは，MLLMs [Achiam 23, GeminiTeam 23, Dai 23] を利用し，自然言語を通じて複雑な指示表現や位置関係を表現する構造的特徴を抽出するものである．本研究では，この種の特徴を Narrative Representation と定義する．

既存の多くの手法は，上記のすべての視覚表現を包括的に扱えておらず，視覚特徴の表現力が制限されている．具体的には，Scene Representation は画像から形状や色などの視覚情報を捉えることができるものの，位置関係を含む複雑な参照関係を抽出することは困難であり，この表現だけでは不十分である．また，Narrative Representation は自然言語を通じて構造的な特徴を抽出できるが，テクスチャなどの詳細な視覚特徴を十分に捉えることは難しい．一方，Aligned Representation は自然言語とアラインされており，Scene Representation と Narrative Representation の両方の特性を兼ね備えている．しかし，Aligned Representation のみを用いる場合，指示文中の複雑な参照表現を構造的に理解する能力が不足しやすい．これは，この表現が自然言語を通じた構造的な特徴抽出を行わないためである．以上のことから，これらの視覚表現をすべて並行して活用することで，十分な視覚表現を得ることができると考えられる．

3.2 λ-Representation Encoder

本研究では，λ-Representation を生成する λ-Representation Encoder を導入する．本モジュールでは，3 種類の視覚表現を取得し，それらを統合することで λ-Representation を生成する．図 1 に示すように，このモジュールは Scene Representation Module，Aligned Representation Module，Narrative Representation Module の 3 つのサブモジュールで構成されている．λ-Representation Encoder の入力は $\boldsymbol{x}_{\text{before}}$ または $\boldsymbol{x}_{\text{after}}$ のいずれかである．なお，$\boldsymbol{x}_{\text{after}}$ に対しても同様の処理を行うため，以下の説明では $\boldsymbol{x}_{\text{before}}$ に対して行う操作を説明する．

まず，Scene Representation Module を $f_{\text{srm}}(\cdot)$ としたとき，Scene Representation の取得は $\boldsymbol{h}_s = f_{\text{srm}}(\boldsymbol{x}_{\text{before}})$ と表すことができる．Scene Representation Module は，複数のバックボーンネットワークで構成されている．本研究では，バックボーンネットワークとして ViT [Dosovitskiy 20]，DINOv2 [Darcet 24]，CLIP 画像エンコーダ [Radford 21] を使用する．ViT および DINOv2 については，出力をそのまま用い，CLIP の画像エンコーダは中間特徴を使用する．最終的に，これらの特徴を連結することで，$\boldsymbol{h}_s$ を得る．

次に，Aligned Representation Module を用いて Aligned Representation $\boldsymbol{h}_a$ を取得する．このモジュールは，マルチモーダル基盤モデルで構成されている．これらの特徴量は，自然言語と適切にアラインされているため，Aligned Representation と解釈できる．本研究では，CLIP の画像エンコーダを用い，その出力を抽出する．

最後に，Narrative Representation Module を用いて，Narrative Representation $\boldsymbol{h}_n$ を取得する．このモジュールは，MLLM および複数のテキストエンコーダで構成されている．本研究では，$\boldsymbol{x}_{\text{before}}$ から画像のキャプションを生成するために InstructBLIP [Dai 23] を使用する．キャプション生成の際，物体の色，大きさ，形状に加え，それらの配置，画像内での位置関係，および他の物体との相対的な位置関係に注目するようなテキストプロンプトを設計した．InstructBLIP の出力は，BERT および text-embedding-3-large [OpenAI 24] を用いて特徴を抽出し，これらを連結することで $\boldsymbol{h}n$ を得る．最終的に，$\boldsymbol{x}_{\text{before}}$ の λ-Representation $\boldsymbol{h}_\lambda = \left[\boldsymbol{h}_s^\top, \boldsymbol{h}_a^\top, \boldsymbol{h}_n^\top\right]^\top$ を取得する．同様に，$\boldsymbol{x}_{\text{after}}$ の λ-Representation を $\boldsymbol{h}'_\lambda$ とする．

3.3 Contrastive λ-Representation Decoder

本研究では，$\boldsymbol{h}_\lambda$ と $\boldsymbol{h}'_\lambda$ の差異を表現するために，Contrastive λ-Representation Decoder を導入する．物体操作の影響は画像間の変化に含まれているため，本表現を用いることで，モデルは操作に起因する差分に着目することが可能となる．一方で，画像間の変化が，指示文で指定されたタスクの成功を示すとは限らない．例えば，図 1 に示されているケースでは，Pepsi の缶が倒れた場合，画像間に差異が生じるが，この操作

手法	Accuracy [%]
Xiao et al. [Xiao 22]	71.59 ± 1.95
Contrastive λ-Repformer	**80.80** ± 0.86

表 1: SP-RT-1 データセットにおける Xiao らの手法および Contrastive λ-Repformer の定量的結果．最も精度が高かった手法の精度を太字で示している．

は失敗とみなされるべきである．したがって，画像間の差異のみをもとに操作の成否を判断するのは困難である．そのため，物体操作の成功または失敗を予測する際には，差異表現と指示文とのアラインメントを考慮することが重要である．

本モジュールの入力は $\boldsymbol{h}_\lambda$，$\boldsymbol{h}'_\lambda$，$\boldsymbol{h}_l$ であり，出力は $P(\hat{y}=1)$ である．まず，2 つの画像の差異を表現する特徴量 $\boldsymbol{h}_{\text{diff}}$ を以下のように取得する．

$$\boldsymbol{h}_{\text{diff}} = \text{CrossAttn}\left(\boldsymbol{h}'_\lambda, \boldsymbol{h}_\lambda\right), \tag{1}$$

ただし，$\text{CrossAttn}(\cdot,\cdot)$ は cross-attention 機構を表す．cross-attention 機構は，任意の行列 $\boldsymbol{X}_A$ および $\boldsymbol{X}_B$ に対して，以下のように定義される．

$$\text{CrossAttn}(\boldsymbol{X}_A, \boldsymbol{X}_B) = \text{softmax}\left(\frac{\boldsymbol{X}_A\boldsymbol{W}_q(\boldsymbol{X}_B\boldsymbol{W}_k)^\top}{\sqrt{d_k}}\right)\boldsymbol{X}_B\boldsymbol{W}_v,$$

ただし，$\boldsymbol{W}_q$，$\boldsymbol{W}_k$，$\boldsymbol{W}_v$ は学習可能な重みであり，d_k は $\boldsymbol{X}_B\boldsymbol{W}_k$ の次元を表す．また，$\boldsymbol{h}_{\text{diff}}$ と $\boldsymbol{h}_l$ の間のアラインメント表現 $\boldsymbol{h}_{\text{align}}$ を以下のように得る．

$$\boldsymbol{h}_{\text{align}} = \text{CrossAttn}(\boldsymbol{h}_{\text{diff}}, \boldsymbol{h}_l) \tag{2}$$

最後に，このモジュールの出力として，多層パーセプトロンを用いて $\boldsymbol{h}_{\text{align}}$ から $P(\hat{y}=1)$ を得る．損失関数はクロスエントロピー損失を使用する．

4. 実験設定

実験では RT-1 データセット [Brohan 22] および SP-RT-1 データセット [齋藤 24] を利用した．本タスクはそれぞれのエピソードにおいて指示文，物体操作時に撮影された画像，および物体操作の成否を示すラベルが必要である．まず，RT-1 データセットは実世界での open-vocabulary 物体操作のための標準的な大規模データセットである．このデータセットは指示文，物体操作中に収集された画像，および二値報酬を含む．よって，このデータセットを用いてリアルタイム物体操作タスク成功判定での評価を行う．RT-1 データセットを直接 SPOM タスクに用いることはできないため，RT-1 データセットから作成されたデータセットが SP-RT-1 データセットである．

また，ベースライン手法を Xiao らの手法 [Xiao 22] とした．この手法は 2 つの画像と指示文に基づく失敗検出モデルであり，ロボットによる物体操作向けの大規模モデルである PaLM-E [Driess 23] と同等であると報告されているため，ベースライン手法として選択した．

5. 実験結果

5.1 定量的結果

表 1 は Contrastive λ-Repformer と Xiao らの手法の定量的結果を示す．SP-RT-1 データセットにおいて，Contrastive λ-Repformer の精度は 80.80%であり，精度が 71.59%であった Xiao らの手法を 9.21 ポイント上回った．この結果より，提案手法は従来手法よりも優れた性能を示すことを示唆している．

モデル	Freeze	SR			AR	NR	Acc. [%]
		CLIP	ViT	DINOv2			
(i)	✓	✓	✓	✓	✓	✓	**80.8**
(ii)		✓	✓	✓	✓	✓	79.2
(iii)					✓	✓	73.7
(iv)		✓			✓	✓	67.7
(v)			✓		✓	✓	77.5
(vi)				✓	✓	✓	79.9
(vii)		✓	✓	✓		✓	77.1
(viii)		✓	✓	✓	✓		75.2

表 2: SP-RT-1 データセットにおける ablation study の結果．最も精度が高かったモデルの精度を太字で示している．本表において，freeze，SR，AR，NR はそれぞれ，バックボーンネットワークのパラメータの freeze，Scene Representation，Align Representation，Narrative Representation を表す．

5.2 Ablation Study

バックボーンネットワークのパラメータを freeze することによる精度への影響を調査するため，それらの unfreeze に関する ablation study を行った．表 2 は ablation study の結果を示している．表に示されているように，SP-RT-1 データセットにおいて，バックボーンネットワークを unfreeze したモデルの精度は，すべてのバックボーンネットワークを freeze して使用したモデル (i) のものと比較して低かった．一方で，バックボーンネットワークを unfreeze した場合，モデル (vi) はモデル (ii) よりも精度が 0.7 ポイント高かった．これは，バックボーンネットワークのパラメータを unfreeze したモデル間の比較において，単純なアーキテクチャがより効果的となる場合があることを示している．しかし，バックボーンネットワークのパラメータを freeze し，すべてのバックボーンネットワークを活用する提案モデルの精度が最も高いことが確認された．

5.3 動画のタスク成功判定への適用

Contrastive λ-Repformer を動画のタスク成功判定に適用した．本手法は，SPOM タスクを実行する際に 2 枚の画像のみを入力として扱うが，この手法を用いて動画に基づく成否判定を実行することが可能である．本タスクは，入力画像対に対して，各時刻における物体操作の成否を $(t=0, t=1), (t=0, t=2), \ldots, (t=0, t=N-1), (t=0, t=N)$ のように予測することで実行可能である．ただし，$(t=0, t=n)$ は時刻 $t=0$ と $t=n$ におけるフレームからなる画像対を表す．このフレームワークでは，出力が任意の時刻で ‘Success’ となるか，最後まで ‘Failure’ を出力し続けるかにもとづいて，動画のタスク成功判定を行うことができる．

図 2 に成功例を示す．このサンプルの指示文は “pick green rice chip bag”，正解ラベルは ‘Success’ であった．また，サンプルには 16 フレームあり，$t=14$ のときに出力結果が ‘Failure’ から ‘Success’ に変わった．つまり，提案手法はこの変化を適切に検出することができた．これは，Contrastive λ-Repformer が動画のタスク成功判定にも適用可能であることを示している．本手法の利点は，動画全体を入力とする手法とは異なり，リアルタイムでの成功判定が可能である点にある．

6. おわりに

本研究では，指示文と物体操作の前後に撮影された一人称視点画像が与えられたときに，open-vocabulary 物体操作の成否を予測するタスクを扱った．本研究の貢献は以下の通りである．λ-Representation Encoder を導入し，多階層アラインメント視覚表現 λ-Representation を生成した．この表現は，色や形

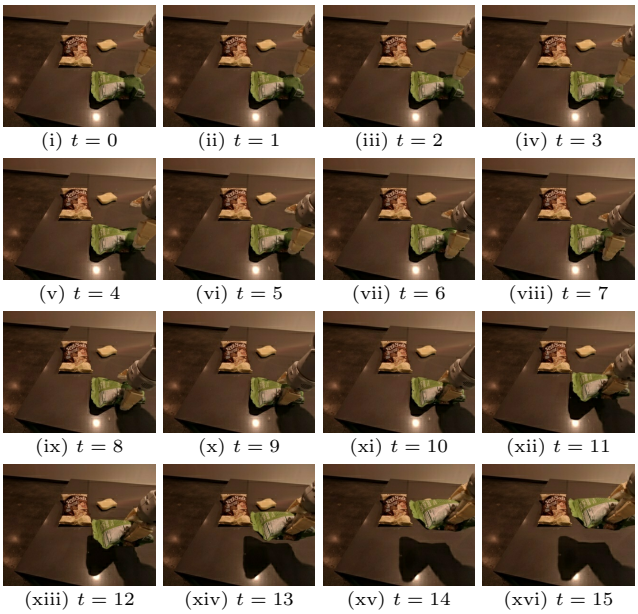

図 2: Contrastive λ-Repformer を用いた動画のタスク成功判定における成功例．画像は，時刻 $t = 0$ から $t = 15$ のフレームに対応する．本サンプルは，RT-1 データセットから取得したエピソードである．本サンプルにおいて，ロボットマニピュレータに与えられた指示文は “pick green rice chip bag”，正解ラベルは ‘success’ であった．

状などの視覚的特徴を保持する特徴，自然言語とアラインされた特徴，自然言語を通じて構造化された特徴の 3 つの要素から構成される．また，Contrastive λ-Representation Decoder を導入した．本モジュールにより，2 つの画像間の差異を抽出し，その差異と指示文のアラインメントを考慮することが可能となった．さらに，Contrastive λ-Repformer がリアルタイム物体操作タスク成功判定に適用可能であることを確認した．将来研究として，本手法をより広範な物体操作およびナビゲーションタスクへ適用することが挙げられる．

謝辞

本研究の一部は，JSPS 科研費 23K03478，JST ムーンショット，JSPS 特別研究員奨励費 JP23KJ1917 の助成を受けて実施されたものである．

参考文献

[Achiam 23] Achiam, J., Adler, S., Agarwal, S., et al.: GPT-4 Technical Report, *arXiv preprint arXiv:2303.08774* (2023)

[Brohan 22] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., et al.: RT-1: Robotics Transformer for Real-World Control at Scale, *arXiv preprint arXiv:2212.06817* (2022)

[Chen 20] Chen, Y., Li, L., Yu, L., Kholy, E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J.: UNITER: UNiversal Image-TExt Representation Learning, in *ECCV*, pp. 104–120 (2020)

[Dai 23] Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S.: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, in *NeurIPS* (2023)

[Darcet 24] Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P.: Vision Transformers Need Registers, in *ICLR* (2024)

[Dosovitskiy 20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in *ICLR*, pp. 12888–12900 (2020)

[Driess 23] Driess, D., Xia, F., Sajjadi, M., Lynch, C., Chowdhery, A., Ichter, B., et al.: PaLM-E: An Embodied Multimodal Language Model, in *ICML*, Vol. 202, pp. 8469–8488 (2023)

[GeminiTeam 23] GeminiTeam, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, B., Yu, J., Soricut, R., Schalkwyk, J., et al.: Gemini: A Family of Highly Capable Multimodal Models, *arXiv preprint arXiv:2312.11805* (2023)

[Goko 24] Goko, M., Kambara, M., Saito, D., Otsuki, S., and Sugiura, K.: Task Success Prediction for Open-Vocabulary Manipulation Based on Multi-Level Aligned Representations, in *CoRL*, pp. 3242–3263 (2024)

[Li 22] Li, J., Li, D., Xiong, C., and Hoi, S.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, in *ICML*, pp. 12888–12900 (2022)

[Liu 21] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, in *ICCV*, pp. 10012–10022 (2021)

[OpenAI 24] OpenAI, : text-embedding-3-large (2024), Accessed: May. 2024

[Radford 21] Radford, A., Wook, K., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., et al.: Learning Transferable Visual Models From Natural Language Supervision, in *ICML*, pp. 8748–8763 (2021)

[Xiao 22] Xiao, T., Chan, H., Sermanet, P., Wahid, A., Brohan, A., Hausman, K., Levine, S., and Tompson, J.: Skill Acquisition by Instruction Augmentation on Offline Datasets, in *LangRob @ CoRL22* (2022)

[Zhai 23] Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L.: Sigmoid Loss for Language Image Pre-Training, in *ICCV*, pp. 11975–11986 (2023)

[齋藤 24] 齋藤 大地, 神原 元就, 九曜 克之, 杉浦 孔明：マルチモーダル LLM および視覚言語基盤モデルに基づく大規模物体操作データセットにおけるタスク成功判定, 第 38 回人工知能学会全国大会資料, pp. 3O1OS16b02–3O1OS16b02 (2024)